ARMSTRONG LABORATORY

TREATMENT OF OUTLIERS IN COGNITIVE AND PSYCHOMOTOR TEST DATA

Charles E. Lance
Amy M. Stewart

Department of Psychology
University of Georgia
Athens, GA 30602

Thomas R. Carretta

HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352

March 1993

Interim Technical Paper for Period January 1992 - October 1992

Approved for public release; distribution is unlimited.

AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS

NOTICES

This technical paper is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation, or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

Thomas R. Carretta
THOMAS R. CARRETTA
Project Scientist
Manpower & Personnel Research Div

William E. Alley
WILLIAM E. ALLEY
Technical Director
Manpower & Personnel Research Div

W. Alford
ROGER W. ALFORD, Lt Colonel, USAF
Chief, Manpower & Personnel Research Div

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
March 1993

3. REPORT TYPE AND DATES COVERED
Interim - January 1992 - October 1992

4. TITLE AND SUBTITLE
Treatment of Outliers in Cognitive and Psychomotor Test Data

5. FUNDING NUMBERS

C	F41689-88-D-0251
PE	62205F
PR	7719
TA	18
WU	54

6. AUTHOR(S)

Charles E. Lance
Amy M. Stewart
Thomas R. Carretta

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Department of Psychology
University of Georgia
Athens, GA 30602

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

Armstrong Laboratory
Human Resources Directorate
Manpower and Personnel Research Division
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AL-TP-1993-0002

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Many statistical tests and review articles have pointed out the possible adverse effects that outliers can have on model parameter estimates, and have suggested several methods for detecting and treating outliers. In the present study, the effects of two different methods for treating outliers in aptitude tests (data deletion and data transformation) were investigated at the item and total-score level on the internal consistency and criterion-related validity of six computerized tests being evaluated by the U.S. Air Force. Over 2,000 pilot training candidates were tested. Results indicated that neither outlier treatment method at either level of analysis had significant effects on tests' psychometric characteristics. Possible reasons for these findings include the rarity with which outliers actually occur, and the robustness of linear modeling methods.

14. SUBJECT TERMS

Computer-based tests	Test reliability
Outliers analysis	Test validity

15. NUMBER OF PAGES
20

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

CONTENTS

TABLES

PREFACE

This research and development effort was accomplished under Work Unit 77191854, Contract Number F41689-88-D-0251, QY27, Task 57 in support of aircrew selection and classification. The authors thank Drs Malcolm James Ree, Joseph L. Weeks, and William E. Alley for their many helpful comments and suggestions.

Treatment of Outliers in Cognitive and Psychomotor Test Data

SUMMARY

Many statistical tests and review articles have pointed out the possible adverse effects that outliers can have on model parameter estimates, and have suggested several methods for detecting and treating outliers. In the present study, the effects of two different methods for treating outliers in aptitude tests (data deletion and data transformation) were investigated at the item and total-score level on the internal consistency and criterion-related validity of six computerized tests being evaluated by the U.S. Air Force. Over 2,000 pilot training candidates were tested. Results indicated that neither outlier treatment method at either level of analysis had significant effects on tests' psychometric characteristics. Possible reasons for these findings include the rarity with which outliers actually occur, and the robustness of linear modeling methods.

INTRODUCTION

Data points which lie apart from the majority of the data, or outliers, can have a significant impact on model parameter estimates (Chatterjee & Hadi, 1986; Cook & Weisberg, 1982; Maier, 1988; Neter, Wasserman, & Kutner, 1990; Stevens, 1984). There may be many sources of outlying data points including (a) collecting data from subjects (e.g., pre-teens) who are not members of targeted population (adults), (b) extreme contributions of random error components, (c) data recording errors, and (d) errors in data preparation (Orr, Sackett, & Dubois, 1991).

Although several methods for detecting and treating outliers have been developed (see Belsley, Kuh, & Welsh, 1980 and Chatterjee & Hadi, 1986 for reviews), issues of outlier detection and treatment have received scant attention in the human resource management literature. As one example, Orr et al. (1991) reviewed 100 selection validation studies cited by Schmitt, Gooding, Noe, and Kirsh (1984) which had been published between 1964 and 1982 in *Personnel Psychology* and the *Journal of Applied Psychology*. According to Orr et al. (1991), "not a single study mentioned looking for, finding, or removing outlying data" (p. 475). This is surprising, given the influence that a small number, or even one, data point could have on model parameter estimates (Cook & Weisberg, 1982).

To examine possible effects of outliers on test validities, Orr et al. (1991) investigated the effects of outlier removal on the validity of selected General Aptitude Test Battery (GATB; see Hunter, 1980 for a description of the GATB) tests in a set of 183 studies. They found that (a) "outlying data points were not...a substantial source of variance" (p. 473), (b) removing outliers had little effect on mean validities, and (c) removing outlying data points often reduced, rather than increased, test validities. There are at least three explanations for these somewhat unexpected findings. First, deleted observations may have been extreme but not outlying. Thus, the deleted observations may not have been outliers but well behaved data points instead. Second, deleting observations lying in the tails of the score distribution may have artifactually restricted the range of the score variance and consequently reduced validity. Finally, the type of paper-and-pencil scores studied by Orr et al. (1991) may not be as sensitive to the threat of outliers as other types of scores (e.g., response latencies, tracking error). Paper-and-pencil tests typically are scored in terms of percentages (e.g., percent correct, percent completed), and may be converted to some other metric (e.g., z-scores, percentiles) for interpretive ease, so that out-of-range values are easily recognized and corrected, and scores are bounded by admissible values (e.g., 0% to 100% correct, 1st through the 99th percentile).

Outliers may be more influential in other types of measures such as response latencies and/or measures of psychomotor performance. Response latencies characteristically are positively skewed and often contain outlying "long" responses (Luce, 1986; Teichner & Krebs, 1972, 1974). This becomes an issue because along with simple adaptation of paper-and-pencil tests to computer administration, computerized test scoring allows the measurement of additional dimensions of performance beyond simple accuracy scores (correct/incorrect) such as response latency and tracking error. Thus, while outliers may not be pervasive and influential in traditional paper-and-pencil tests, they might be more so on computerized tests, and particularly on performance measures which characteristically are nonnormal (Green, 1988).

There are two further issues with respect to outliers in response latency and tracking error data which ordinarily are not issues for tests containing items scored correct/incorrect: (a) level of analysis, and (b) whether data should be deleted or transformed to treat outliers. First, outliers simply do not exist at the item level for tests scored correct/incorrect. An individual's response is either right, or it is wrong. For these tests, outliers exist only at the total score level. Outlying respondents (a) answer a large proportion of items correctly or incorrectly or (b) skip or do not attempt many items. On the other hand, outlying response latencies can be extremely short (e.g., an anticipatory response), or extremely long on individual test items. Similarly, tracking errors can be very small or very large for particular scoring intervals (i.e., time segments) as well as the test overall. Thus, the issue is whether outlying responses should be treated at the item level, or only at the total score level, given that a sufficient number of item-level outliers cumulate to produce an outlying total test score.

The second issue is whether outlying data, either at the item- or total-score level should be deleted, or whether the data should be transformed so that possible ill effects of outlying data points on model parameter estimates are reduced. Much of the literature on outliers and influential data points has focused on how to identify data points to delete (e.g., Chatterjee & Hadi, 1986; Cook & Weisberg, 1982), with little or no attention paid to the effects of data transformations. On the other hand, researchers accustomed to working with reaction time data routinely effect logarithm transformations to more nearly normalize the data, and rarely delete data points (Luce, 1986).

Purpose

Computerized testing permits the measurement of additional dimensions of performance beyond those measured by traditional paper-and-pencil tests such as response latency and psychomotor performance. However, the availability of these measures raises issues of how the data should be treated psychometrically. This study focused on the treatment of outlying data on a representative computerized test battery (Basic Attributes Test or BAT; see Carretta, 1990).

Measures

The BAT battery used in this study consisted of six computerized tests that assessed individual differences in psychomotor coordination, (rotary pursuit, compensatory tracking), information processing ability (spatial transformation, short-term memory, and time sharing ability), and attitudes toward risk. The types of scores generated from these tests include tracking error, response time, response accuracy, and response choice. Several studies have shown that BAT scores are useful for predicting US Air Force pilot training performance and provide incremental validity when used with operational selection instruments such as the AFOQT (Bordelon & Kantor, 1986; Carretta, 1989; Kantor & Carretta, 1988). Operational implementation of the BAT as an adjunct to current pilot candidate selection methods is expected in 1993 (Carretta, 1992). A brief description of the BAT selection tests follows; a more detailed description was provided by Carretta (1989, 1990).

Two-Hand Coordination. This pursuit tracking task was used to measure multilimb coordination (Fleishman, 1964). An airplane (target) moved in a fixed, elliptical pattern at a varying rate. The subject controlled the horizontal and vertical movement of a "gunsight" using the right (horizontal) and left (vertical) control sticks. The subject's task was to keep the gunsight on the target. Horizontal and vertical tracking error was scored for each of ten, 30 second intervals (n = 2,451).

Complex Coordination. This compensatory tracking task was an example of control precision and multilimb coordination (Fleishman, 1964). The dual-axis right control stick was used to control the horizontal and vertical movement of a cursor. The left control stick was used to control the left-right movement of a "rudder bar" at the base of the screen. The subject's task was to maintain the cursor (against a constant horizontal and vertical rate bias) centered on a large cross fixed at the center of the screen, while simultaneously centering the rudder bar at the base of the screen (also, against a constant rate bias). Horizontal and vertical tracking error was scored for each of ten, 30 second intervals (n = 2,451).

Mental Rotation. This was a variation of a spatial transformation task (Shepard & Metzler, 1971). The subject was presented sequentially with two letters and was required to make a same-different judgment. The letter pair consisted of either same or minor images and the letters were either in the same orientation or rotated in relation to each other. A correct "different" judgment is associated with a mirror image pair and is not dependent on the relative rotation of the two letters. Response speed and accuracy were scored for each of the 72 items for this test (n = 2,147).

Item Recognition. This measure of short-term memory was based on a task proposed by Sternberg (1966). A string of 1 to 6 digits was presented on the screen. The string was then removed, and after a brief delay, replaced by a single digit. The subject was instructed to remember the digit string and determine whether the single digit was one of those presented in the digit string. Response speed and accuracy were scored for each of the 48 items for this test (n = 2,209).

Time Sharing. This test provided a measure of time sharing performance (North & Gopher, 1976). In the first 10 minutes of this test, the subject was required to keep a randomly moving "gunsight" or an airplane (target) using the right-hand control stick. In the next six minutes, the subject had to repeat the tracking task and simultaneously cancel digits which appeared at random intervals and locations on the screen. Digit cancellation was timed and consisted of pressing the same digit on the numeric keypad. The final three minutes of the test consisted of only the tracking task. Tracking difficulty was varied by increasing or decreasing the control stick sensitivity as a function of the tracking error. Scores for this test included tracking difficulty and response time (n = 2,356).

Activities Interest Inventory. This test was designed to measure the subject's attitudes toward risk-taking (Mullins, 1962). The subject was presented with 81 pairs of activities and was asked to choose between them. The activity pairs forced the subject to choose between activities that differed as to degree of threat (sometimes subtly, sometimes not). Response speed and response choice were scored for each item (n = 2,355).

Apparatus

The test apparatus consists of a microcomputer and monitor built into a ruggeoized chassis with a glare shield and side panels designed to minimize distractions. The subjects responded to the tests by manipulating individually or in combination, a dual-axis control stick on the right side, a single-axis control stick on the left side, and a keypad in the center of the test unit. The keypad included keys labeled 0 to 9, an ENABLE key in the center, and a bottom row with YES and NO keys, and two others for same/different responses (S/D), and left/right responses (L/R).

Procedure

All subjects were enrolled in a 4-year college program. They were tested on the AFOQT either prior to entering college or while an undergraduate and were tested on the BAT battery at the beginning of a flight screening program prior to receiving flying training.

The tests used in this study were part of a longer test battery that required about four hours to complete including programmed breaks between the tests. After the test administrator briefed the subjects and initialized the test battery, the test session was self-paced by each subject. Programmed breaks of one or two minutes between tests were included in order to reduce mental and physical fatigue.

Training Performance

Undergraduate Pilot Training (UPT) is a 53 week program involving a T-37 phase (initial jet trainer, 21 weeks) and a T-38 phase (advanced jet trainer, 32 weeks). UPT final outcome was awarded at the end of the program and was scored as a dichotomous variable. Graduates received a score of 1 and eliminees received a score of 0.

Identification and Deletion of Outliers

Although there are numerous methods for detecting outlying data points (e.g., Chatterjee & Hadi, 1986), graphical methods still are among the most effective. We used univariate frequency distributions to identify outliers both at the item- and total-score level. Specifically, they were examined for points at which there were discontinuities in the frequency distributions, beyond which very few data points lay. In most cases, this corresponded closely to the 1st and/or 99th percentiles of the distribution. Outlying data points were defined to lie beyond these points. Sets of total scores were created both with outlying item scores included and deleted. Also, total scores that included outlying item scores were themselves examined for outlying values and were censored for outliers at the total score level.

Data Transformations

Tracking error and response latency data were markedly positively skewed. For this type of distribution, Mosteller and Tukey (1977) recommend either a square-root or a natural log transformation to more closely normalize the data. We effected both transformations at both the item- and total-score level. Thus, two sets of total scores were created. The first was based on nontransformed, square root transformed, or log transformed item-level data, that were summed to form a total score (item-level). The second was based on nontransformed item-level data that were summed to form a total score and then remained nontransformed or was square root or log transformed (total score).

Analysis

Internal consistency reliability (Cronbach's alpha [Cronbach, 1951]) was estimated for each BAT score for both nontransformed and transformed data. Cronbach's alpha is the most widely cited measure of internal consistency. It is the average of all split-half reliability coefficients, a measure of test homogeneity, and an estimate of first factor saturation (Stanley, 1971). Test score validities were estimated by correlating nontransformed and transformed BAT summary scores with UPT final outcome (graduation or elimination).

RESULTS

Between 0.27% (Item Recognition response time) and 15.65% (Mental Rotation response time) of the subjects were identified as having outlying values at the item-level for nontransformed response latency or tracking performance measures. The mean percent of subjects having outlying item-level data was 3.98% for these scores. The proportion of subjects with outlying values at the total-score level for nontransformed data ranged from 0.13% (Activities Interest Inventory average response time) to 2.23% (Complex Coordination, horizontal tracking error) with a mean of 0.76%.

The treatment of outliers at the item-level (inclusion or exclusion; nontransformed or transformed) had little effect on the internal consistency estimates. Internal consistency estimates (Cronbach's alpha), shown in Table 1, were acceptably high for all scores. The mean internal consistency ranged from .929 to .947 across the outlier treatment methods.

Correlations between UPT final outcome and test scores based on item level and total-score level data are summarized in Tables 2 and 3.

Neither deleting outliers nor transforming data at the item- or total-score level had much impact on the test score correlations with UPT final outcome. In the case of the test scores based on item-level data (Table 2), outlier deletion and data transformation actually lowered many validity coefficients.

DISCUSSION

In particular, it is noteworthy that the inclusion or removal of outliers had little influence on the internal consistency and predictive validity of the BAT scores. This is important as these tests are expected to become operational US Air Force pilot candidate selection instruments in the near future (Carretta, 1992). Under operational conditions, it will be desirable for all pilot training applicants to receive meaningful test scores so that personnel selection decisions can be made.

Table 1. Coefficient Alpha for Nontransformed and Transformed Scores

Score	N Items	Nontransformed Data		SQRT Transformed		Log Transformed	
		W/Outliers	W/O Outliers	W/Outliers	W/O Outliers	W/Outlier	W/O Outliers
Two-Hand Coordination							
Horizontal Error	10	.99	.99	.99	.98	.99	.98
Vertical Error	10	.99	.98	.99	.99	.99	.99
Complex Coordination							
Horizontal Error	10	.96	.95	.97	.96	.97	.97
Vertical Error	10	.96	.94	.97	.96	.97	.97
Rudder Error	10	.95	.93	.96	.95	.96	.95
Mental Rotation							
Response Time	72	.98	.98	.99	.98	.99	.99
Response Outcome	72	.95	.94	.95	.95	.95	.94
Item Recognition							
Response Time	48	.98	.98	.99	.98	.98	.99
Response Outcome	48	.77	.69	.77	.65	.69	.65
Time Sharing							
Difficulty Level	19	.99	.99	.99	.99	.99	.99
Response Time	35	.88	.86	.90	.88	.89	.89
Activities Interest Inventory							
Response Time	81	.95	.96	.97	.96	.97	.97
Response Choice	81	.87	.88	.87	.88	.87	.88
MEAN		.941	.929	.947	.932	.939	.937

Note. The column labeled "N Items" refers to time intervals for the tests involving tracking performance (Two-Hand Coordination, Complex Coordination, and Time Sharing).

Table 2. Correlation With UPT Final Outcome for Scores Based on Nontransformed and Transformed Data at Item Level

Score	Nontransformed Data		SQRT Transformed		Log Transformed	
	W/Outliers	W/O Outliers	W/Outliers	W/O Outliers	W/Outlier	W/O Outliers
Two-Hand Coordination						
Horizontal Error	.18*	.20*	.20*	.20*	.21*	.22*
Vertical Error	.19*	.18*	.20*	.18*	.20*	.18*
Complex Coordination						
Horizontal Error	.08*	.08*	.08*	.07*	.07*	.06*
Vertical Error	.09*	.10*	.10*	.10*	.11*	.10*
Rudder Error	.11*	.11*	.12*	.11*	.11*	.10*
Mental Rotation						
Response Time	.07*	.07*	.07*	.04	.03	.02
Response Outcome	.04	.04	.03	.02	.03	-.02
Item Recognition						
Response Time	.11*	.11*	.00	.00	.01	.00
Response Outcome	.03	.03	.03	.03	.02	.03
Time Sharing						
Difficulty Level	.08*	.08*	.08*	.08*	.08*	.05*
Response Time	.21*	.20*	.23*	.21*	.24*	.20*
Activities Interest Inventory						
Response Time	.02	.03	.00	.00	.01	.00
Response Choice	.01	.01	.01	.01	.01	.01
MEAN	.094	.095	.089	.081	.087	.076

Note. Correlation signs for tracking error (Two-Hand Coordination, Complex Coordination) and response time scores (Mental Rotation, Item Recognition, Time Sharing, Activities Interest Inventory) were reflected so that higher scores indicate better performance.
*p < .01

Table 3. Correlation With UPT Final Outcome for Scores Based on Nontransformed and Transformed Data at Total-Score Level

Score	Nontransformed Data		SQRT Transformed		Log Transformed	
	W/Outliers	W/O Outliers	W/Outliers	W/O Outliers	W/Outlier	W/O Outliers
Two-Hand Coordination						
Horizontal Error	.18*	.19*	.20*	.21*	.21*	.21*
Vertical Error	.19*	.17*	.20*	.17*	.19*	.17*
Complex Coordination						
Horizontal Error	.08*	.09*	.09*	.09*	.08*	.08*
Vertical Error	.09*	.12*	.10*	.12*	.11*	.12*
Rudder Error	.11*	.12*	.12*	.12*	.12*	.12*
Mental Rotation						
Response Time	.07*	.07*	.08*	.06*	.07*	.06*
Response Outcome	.04	.05*	.03	.04	.03	.04
Item Recognition						
Response Time	.11*	.10*	.11*	.10*	.11*	.10*
Response Outcome	.03	.04	.03	.04	.03	.04
Time Sharing						
Difficulty Level	.08*	.09*	.08*	.09*	.08*	.09*
Response Time	.21*	.22*	.22*	.22*	.22*	.22*
Activities Interest Inventory						
Response Time	.02	.02	.02	.02	.03	.03
Response Choice	.01	.01	.01	.01	.02	.01
MEAN	.094	.099	.099	.099	.099	.099

Note. Correlation signs for tracking error (Two-Hand Coordination, Complex Coordination) and response time scores (Mental Rotation, Item Recognition, Time Sharing, Activities Interest Inventory) were reflected so that higher scores indicate better performance.

*p < .01

Contrary to textbook examples of how extreme data points (outliers) can be unduly influential in model parameter estimates, this is the second study (in addition to Orr et al., 1991) which has shown that in the area of personnel testing, they generally are not. There may be several reasons why.

First, outliers may occur with less frequency than might be expected. For example, Orr et al. (1991) observed that many samples of GATB data failed to contain any cases that qualified as an outlier according to statistical criteria. Thus, the presence of outliers may simply be less of a problem than some have thought. Second, extreme data points may not be outlying as often as they are diagnosed. As an example from the present context, individuals who have extremely long response times to experimental tasks may not make suitable pilots if they also respond very slowly to information received in the cockpit (or worse, they may not be pilots for long!). Third, correlational methods may be robust over attempts to treat outliers. For example, monotonic data transformations such as the square-root or logarithmic affect the shape of the data distribution but do not alter the rank-ordering of the observations.

Results from this study should be interpreted as indicating that outliers do not threaten the integrity of research results, basic or applied. Indeed, both the Orr et al. (1991) study and the present one were conducted in the Federal government under research programs where great care was taken in the collection and preparation of the data bases. Thus, problems of "out of range" data were minimized in both cases. Results do seem to suggest, however, that within carefully constructed data sets, threats of the harmful effects of outliers may not be as serious as some have imagined.

REFERENCES

Belsley, D.A., Kuh, E., & Welsh, R.E. (1980). *Regression diagnostics - identifying influential data and sources of collinearity*. New York: Wiley.

Bordelon, V.P., & Kantor, J.E. (1986). *Utilization of psychomotor screening for USAF pilot candidates: Independent and integrated selection methodologies* (AFHRL-TR-86-4, AD-A170 353). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Carretta, T.R. (1989). USAF pilot selection and classification systems. *Aviation, Space, and Environmental Medicine, 60*, 46-49.

Carretta, T.R. (1990). *Basic Attributes Test (BAT): A preliminary comparison between reserve officer training corps (ROTC) and officer training school (OTS) pilot candidates* (AFHRL-TR-89-50, AD-A224 093). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Carretta, T.R. (1992). Recent developments in U.S. Air Force pilot candidate selection and classification. *Aviation Space and Environmental Medicine, 63*, 1112-1114.

Chatterjee, S., & Hadi, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Review, 1*, 379-416

Cook, R.D., & Weisberg, S. (1982). Characterization of an empirical influence for detecting influential cases in regression. *Technometrics, 22*, 495-508.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Fleishman, E.A. (1964). *The structure and measurement of physical fitness*. Englewood Cliffs, NJ: Prentice Hall.

Green, B.F. (1988). Construct validity of computer-based tests. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 77-86). Hillsdale, NJ: Erlbaum.

Hunter, J.E. (1980). *The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance.* Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Kantor, J.E., & Carretta, T.R. (1988). Aircrew selection systems. *Aviation, Space, and Environmental Medicine, 59*, (11, Supplement), A32-A38.

Luce, R.D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford.

Maier, M.H. (1988). On the need for quality control in validation research. *Personnel Psychology, 41*, 497-502.

Mosteller, F., & Tukey, J.W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Mullins, C.J. (1962). *Objective tests of self-confidence* (PRL-TM-62-66). Lackland AFB, TX: Selection and Classification Branch, Personnel Research Laboratory.

Neter, J., Wasserman, W., & Kutner, M.H. (1990). *Applied linear statistical models* (3rd ed.). Homewood, IL: Irwin.

North, R.A., & Gopher, D. (1976). Measures of attention as predictors of flight performance. *Human Factors, 18*, 1-14.

Orr, J.M., Sackett, P.R., & Dubois, C.L.Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*, 473-486.

Schmitt, N., Gooding, R.Z., Noe, R.A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Shepard, R.N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701-703.

Skinner, J. & Ree, M.J. (1987). *Air force officer qualifying test (AFOQT): Item and factor analyses of form O* (AFHRL-TR-86-68, AD-A184 975). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Stanley, J.C. (1971) Reliability. In R.L. Thorndike (Ed.) *Educational Measurement*, second edition. Washington, DC: American Council on Education.

Sternberg, S. (1966). High speed scanning in human memory. *Science, 153*, 652-654.

Stevens, J.P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*, 334-344.

Teichner, W.H., & Krebs, M.J. (1972). Laws of the simple visual reaction time. *Psychological Review, 79*, 344-358.

Teichner, W.H., & Krebs, M.J. (1974). Laws of visual choice reaction time. *Psychological Review, 81*, 75-98.